Paradigm Development for Identifying and Validating Indicators of Trust in Automation in the Operational Environment of Human Automation Integration

Kim Drnec(✉) and Jason S. Metcalfe

Army Research Laboratory, Aberdeen, MD, USA
kim.a.drnec2.ctr@mail.mil

Abstract. Calibrated trust in an automation is a key factor supporting full integration of the human user into human automation integrated systems. True integration is a requirement if system performance is to meet expectations. Trust in automation (TiA) has been studied using surveys, but thus far no valid, objective indicators of TiA exist. Further, these studies have been conducted in tightly controlled laboratory environments and therefore do not necessarily translate into real world applications that might improve joint system performance. Through a literature review, constraints on an operational paradigm aimed at developing indicators of TiA were established. Our goal in this paper was to develop an operational paradigm designed to develop valid TiA indicators using methods from human factors and cognitive neuroscience. The operational environment chosen was driving automation because most adults are familiar with the task and its consequent structure and therefore required little training. Initial behavioral and survey data confirm that the design constraints were met. We therefore believe that our paradigm provides a valid means of performing operational experiments aimed at further understanding TiA and its psychophysiological underpinnings.

Keywords: Trust in automation · Operational paradigm · Driving automation · Human automation integrated systems

1 Introduction

Joint human automation systems have been developed to leverage the abilities of both agents in order to improve overall task performance. However, true integration has yet to be realized, and the automated agent is often either misused, or disused entirely resulting in relatively poor performance outcomes. One reason genuine integration has not yet been achieved is an apparent lack of user acceptance. The degree to which a human user accepts an automated agent is thought to be directly related to the level of trust the human user has in the automation [1–3]. That is, as people gain confidence in the reliability, robustness, and safety of automated technologies, they develop sufficient trust to willingly share important decision and/or control authority with such systems. Therefore, if automated systems are to be used as designed, enabling joint system performance to reach intended levels, it is important that the human user develop a certain level of trust in the automation (TiA). However, more important than achieving

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC 2016, Part II, LNAI 9744, pp. 157–167, 2016.
DOI: 10.1007/978-3-319-39952-2_16

a certain level of TiA is to manage it so that behavioral outcomes such as misuse or disuse [6–9] do not occur regularly and negatively impact overall performance. Consequently, an important goal for systems designers is to find a means to calibrate the human user's TiA to elicit desired interaction with the automation given the nature of the ongoing, and dynamic, task context [3, 10–13]. An immediate need if TiA is to be calibrated is to establish quantitative and easily monitored indicators of TiA that are robust across individuals, task and time. Currently the only method for assessing TiA is by participant self-report through survey instruments, and few of these surveys have been validated. However, if objective real-time measurements of TiA can be identified and demonstrated as valid, systems could be designed to measure and manage TiA for real world applications that would maximize joint system performance of critical human automation system tasks.

The goal of this paper is to discuss the conceptual underpinnings of an operational research paradigm aimed at inference and validation of TiA outcome measures. First, we discuss important design constraints to such a paradigm based on human factors research. We then provide an example of how these concepts were realized in operational research that adapts methods from cognitive neuroscience and human factors engineering for addressing important issues for TiA and its influence on human-automation systems. Finally, we provide preliminary high-level analysis of an instantiation of our proposed paradigm demonstrating that it meets design constraints, and is therefore suitable as a method to identify indicators of TiA.

2 Concepts for Applying Cognitive Neuroscience to the Operational Study of TiA

Although methods from cognitive neuroscience have been applied in experimental settings to adaptive human automation systems that scale or mitigate task demands on the human user [4, 5] there has been little consistency in how the methods have been applied to specifically study TiA. We propose that such methods, particularly those based in psychophysiology, have considerable potential to effectively identify indicators of TiA if applied under appropriate operational constraints. The basis of this proposal is the understanding that trust is a psychological construct and therefore it would seem reasonable that there would be dynamic psychophysiological variables that enable inferences regarding extant levels of TiA for a given human user. Indeed, research on interpersonal trust has revealed measurable physiological changes correlated with changing participant trust and trust based decision making [6]. Therefore we believe that the application of these cognitive-neuroscience methods is promising for the study of TiA. However, much research across these domains (both cognitive neuroscience and human factors) has been laboratory based, leveraging dramatically simplified tasks performed in controlled environments, often using a narrow set of psychophysiological and/or behavioral data. These methods have resulted in important insights about cognitive and behavioral phenomena underlying human-automation relationships, but these laboratory-based research findings may be of limited value in more complex operational contexts. This is because they tend

to apply to general populations rather than providing an understanding of how human-automation relationships develop as individuals perform tasks with real-world risks and consequences. New research paradigms are therefore required if an understanding of individual relationships are to be understood and leveraged to measure and manage TiA dynamics for particular operational environments.

3 Design Constraints

In order for research in this domain to be of use in operational settings, it is important that experimental conditions engender, as close as is reasonable, authentic levels of trust in ways reflective of operational influences. The relevant literature suggests three critical design considerations if this goal is to be met, (1) establishing a task-relevant risk and consequence structure, (2) engendering TiA levels as a function of automation reliability, and (3) engendering TiA levels as a function of workload. In addition to the theoretically based design constraints, it is critical, if human automation interaction is to be studied, that the subject be motivated to use the automation in a way that is organic to the operational environment. Moreover, we argue that it is critical to develop and validate that these factors have been successfully implemented if the paradigm is to be useful for more detailed research in the cognitive and neural underpinnings of variations in TiA and TiA-related decisions with regard to interactions with automation.

3.1 Risk and Consequence

Development of TiA requires inducing the perception of task-related risk or consequence to the human user [7, 8]; if consequences are low or irrelevant to the human, levels of TiA fail to be important. Generally speaking, we consider that without risk, trust is irrelevant to decision making. In order to develop a sense of risk and consequence it thus appears necessary to facilitate a sense of personal investment in the task outcome. While there may be multiple ways to achieve this, one of the more common methods in research has been to link performance outcomes with extrinsic rewards. Typically, these rewards are financial because most adults have daily experience with financial motivation or gain. Though not directly applicable to many operational contexts, we chose financial motivation as a proven means of creating the needed senses of task investment and risk. Certainly, given the high cost of vehicle-based incidents, financial concerns tend to be common among real-world drivers as well.

3.2 Engendering TiA as a Function of Automation Reliability

Research has yielded much evidence as to what intrinsic and external factors affect extant levels and dynamic changes of TiA in the operational context of human automation integrated systems [1, 8–10]. In the general case, the degree or level of TiA appears to result from the evaluation of observations against *a priori* expectations about how an automation should behave. Initially, most people would expect a real-world automation to be reliable and to be consistent over time, as well as being able to aid in achieving the task

goal [1, 11]. Thus, with some exceptions [11], most human users will have an *a priori* expectation that the automation will be trustworthy, and therefore the initial level of TiA is likely to be relatively high. Reliability, or the degree to which the human user perceives the automation to be accurately performing tasks for which it was designed has significant effects on TiA levels is especially important at the start of automation use. Subsequently, consistency over time becomes critical to dynamic patterns of TiA levels; human users will continue to use, and even benefit from a slightly unreliable (above 70 % reliable) automation if the errors are predictable and consistent over time [12, 13].

3.3 Engendering TiA as a Function of Workload

Workload has been well established as a key influence on behaviors that have traditionally been attributed to TiA. For instance, under high-workload conditions, some people will choose to use an automation for which they hold low trust simply because some assistance is presumed to be better than none [14, 15]. Conversely, research also suggests that under conditions of low workload, human users tend towards manual operating mode [14, 16], likely because of boredom [16]. Therefore, this and other previous research has clearly demonstrated the interaction between workload and trust, leading to the expectation that the effects on psychophysiological variables for each factor would be difficult to disentangle. An operational paradigm focused on real-world outcomes should thus carefully consider the impact of workload in the design of their study on TiA.

3.4 Motivation

If the automation is never used, TiA levels cannot be established, and further, there is no interaction to observe. However, if the participants are rewarded or otherwise explicitly instructed to use the automation, results may reflect experimental design rather than the influence of TiA. One way of motivating natural interaction with the automation is to introduce automation independent secondary task of high value; the logic underlying this is that an automation that sufficiently handles lower value task elements will free operator resources to handle the higher value task. For instance, while modern driving automations are designed to prevent vehicle-vehicle collisions, not all are as capable of predicting and responding to the sometimes erratic and suddenly changing behavior of pedestrians (and other drivers). Therefore, it would be an appropriate driving strategy to engage a driving automation to manage vehicle control, enabling the human occupant to remain vigilant for pedestrians and similar potential hazards.

4 Implementing Operational Constraints into a Research Paradigm

Consider the example of our recently developed leader-follower driving paradigm during which participants were asked to perform a set of tasks relevant to real-world driving. Driving is a model paradigm for our purposes for several reasons. Driving is a task that many people engage in daily, and therefore little training is needed for subjects to perform

an experimental driving task. In addition, driving automations are becoming increasingly common and consequently people are interacting with automations in a natural way. Therefore, an experimental driving paradigm appears to be an excellent operational context to address our questions regarding TiA and human automation interaction.

4.1 Primary Task and Environment

Participants were instructed to drive a simulated vehicle one full lap around a two-lane course. Task objectives included lane position control and maintenance of a "safe" distance from other vehicles, and particularly the lead vehicle in front of them. Automations with different capabilities were presented in different experimental conditions. For conditions in which the automation was available, participants had the option to enable or disable the automation at any moment. Lateral (wind gusts) and longitudinal (lead vehicle speed changes) perturbations were introduced to further challenge the performance of the driving task. In addition, participants were solely responsible for avoiding collisions, as the automation had no explicit collision avoidance capabilities. Therefore, the chosen automation independent secondary task involved avoiding collisions with frequently-appearing pedestrians by responding to them with button presses on a game controller. Pedestrians appeared approximately once every 6 s, distributed randomly on either side of the road, and 15 % stepped in the vehicle path.

4.2 Risk and Consequence

Risk and consequence were expressly manipulated through use of a game-like scenario where each deviation from task parameters had a preset consequence that was known to the participants. The point structure was chosen to encourage a specific hierarchical economy of decision making that was reflective of the risk structure in the real world. For example, collision with a pedestrian incurred the most severe penalty, whereas an

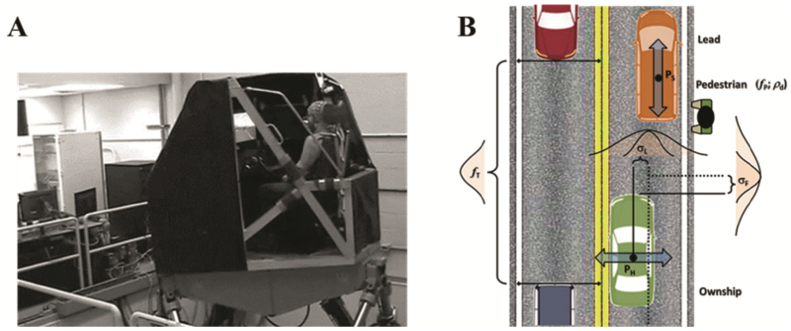

Fig. 1. Summary of experimental paradigm. (A) Ride Motion Simulator shown as a participant completes the driving task while wearing a 64-channel EEG cap. (B) Experimental task. Subjects drove a vehicle (ownship; right lane follow) while following a lead vehicle (right lane lead) and were instructed to maintain following distance and lane position. The varying reliability (low and high) of the driving automation are represented by the distributions labeled σ_L and σ_F.

incorrect button press incurred very little penalty. In order to make the reward significant in the context of adult experience $200 was chosen as a maximum reward, of which $100 could be lost incrementally due to performance decrements. To enhance the realism, and therefore a sense of risk and consequence, participants completed all tasks in an immersive 6-degree of freedom ride motion simulator (Fig. 1A).

4.3 Reliability

In order to develop sensitive measures of TiA it is necessary to encourage a variation of TiA levels both within and across conditions. To this end we implemented two different levels of driving performance reliability; high and low. Reliability characteristics were realized by using lane and speed offsets approximately described by normal distributions with parameters specific to reliability condition as shown in Fig. 1B. The high reliability condition had narrow lane and range offset distributions whereas the distributions in the low reliability automation were broader. The low reliability automation offsets reduced the appearance of consistency over time; here, the offsets were large enough to make it appear that the automation 'wandered' gradually across the lane and following range to varying degrees based on condition.

4.4 Workload

Management of task loading was an important design constraint because of the known interaction between TIA and subjective workload; especially as affecting psychophysiological measures which are of ultimate interest in subsequent analyses. In two "full control" conditions (heading + speed control with low and high reliability) there was an inherent difference in workload owing to reduction in tasking for the human when the automation was performing well. Thus, we expected subjective workload in the high reliability, full (FH) condition to be less than in the low reliability full (FL) condition. The "speed only" conditions were introduced to allow balancing of workload across these conditions. During the speed only, high reliability (SH) condition, it was thought that subjects would primarily need to respond to lateral perturbations because the automation was near perfect in responding to longitudinal perturbations. Conversely, in speed only, low reliability (SL) conditions it would have been necessary to respond to almost all of the perturbations. To balance this circumstance across the speed only conditions, lateral perturbations were introduced more frequently in the SH as compared with the SL, thus aiming to maintain comparable overall workload in both and, importantly, allowing for inferences regarding TIA that were not confounded by effects of increased workload.

4.5 Experimental Design

The average drive time around the course for each condition lasted approximately 12 min. The two different automation capabilities were full control, i.e., both lane and range conforming ability, the second only controlled the speed of the vehicle. Automation reliability groups were high and low reliability. A 2×2 design was realized through automation type (S, F), and automation reliability (L, H); the manual run was treated as

a baseline condition. The experiment consisted of five conditions; manual driving only, full automation with high reliability (FH), full automation with low reliability (FL), speed automation with high reliability (SH), and speed automation with low reliability (SL).

Psychophysiological sensors (electroencephalography (EEG), electrocardiography, galvanic skin response (GSR), and eye tracking) were fitted to each participant and then they completed a 10 min training session where they experienced both types of automation and some of the experimental tasks. After training, data collection began with onset of a manual condition, followed by the other four conditions in a counterbalanced sequence. The course was designed with straight as well as both gradual and sharply curved zones in order to change the likelihood that a trust based decision about automation use would need to be made. Surveys were administered both before the experiment and in between each condition in order to ascertain whether or not we had met our task constraints. The surveys of focus for this paper were the NASA-TLX to assess workload, and trust in automation surveys to gauge TiA levels.

5 Initial Results

Our goal was to develop a paradigm for use in studying TiA as well as neural and cognitive correlates in the operational environment of human automation systems. Previous research aimed at understanding TiA specifies particular, operationally-relevant design constraints that must be met for a successful paradigm to be developed. These include specification of a risk and consequence structure, managing the perceived reliability of automation to influence TiA, and balancing workload. An indication of successful paradigm development, therefore, would be the demonstration of having met these experimental design constraints. Here, we provide subjective survey and behavioral data indicating that our main design was effective.

TiA levels have been shown to be affected by automation reliability. Therefore, it would be expected that low reliability conditions would correspond to low TiA, whereas high reliability conditions would correspond to high TiA. Figure 2A illustrates the relationship between subjective ratings of system trustworthiness and automation reliability by condition. The TiA data were analyzed with a mixed model where reliability and type were fixed and subject data treated as random. There was a significant effect of automation type ($F(1, 71) = 3.47$, $p < 0.05$) and reliability ($F(1, 71) = 71.43$, $p < 0.01$). More important, there was also a significant interaction between automation type and reliability ($F(1, 71) = 5.0$, $p < 0.05$). Figure 2B shows that automation-related decision-making behavior, as revealed in the percentage of time the automation was engaged, reflected the change in apparent TiA as expected.

To examine whether we successfully constructed our paradigm to account for a suspected confound between subjective workload and trust in automation, we assessed the NASA-TLX. Weighted scores are shown in Fig. 3 and hypothesis tests with mixed-model ANOVA confirmed a significant automation type by reliability interaction ($F(1, 71) = 7.8$, $p < 0.05$).

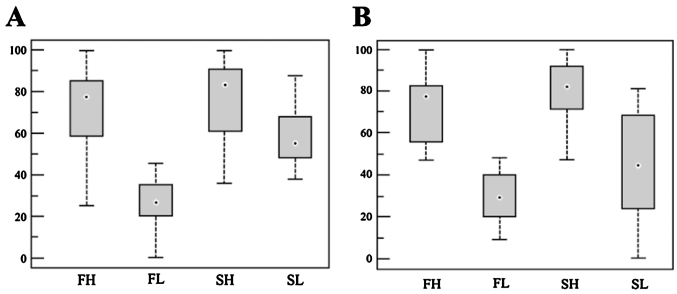

Fig. 2. (A) Subjective ratings (percent) of system trustworthiness, assessed with a visual analogue scale based on Muir (1996) and (B) percent of time automation was used when available.

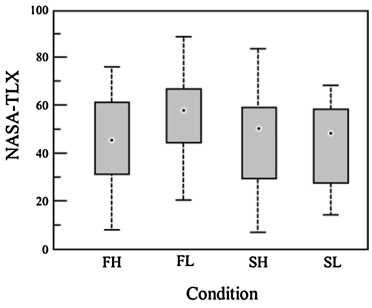

Fig. 3. Overall weighted average scores from the NASA TLX administered at the completion of each driving condition.

6 Discussion

Our aim was to develop an experimental paradigm that allows the study of TiA in the context of interactions with driving automation, an increasingly common operational environment. Behavioral and survey results indicate that we met the required design constraints derived from the TiA literature. For example, TiA levels had a clear relationship with automation reliability conditions, a key factor in TiA development. Figure 2A may also highlight the importance of predictability in TiA preservation; while SL was less trusted than SH, the SL condition appeared to be more trustworthy than the FL condition. This finding likely speaks to the issue of intersecting risk, trust, and predictability. That is, the speed control was likely experienced as generally lower risk than full control because it did not have the capability of steering into the path of an oncoming vehicle. Moreover, its following ability was so *consistently* poor in the SL condition that subjects almost always took over control immediately upon experiencing a longitudinal perturbation. Time spent using the automation should reflect TiA levels. Figure 2B shows the distribution of the percentage of time the automation mode was engaged per condition. One important variable, workload, needed to be controlled for across the speed only conditions. This was done by increasing the number of lateral perturbations that were introduced during the SH condition. Figure 3 gives NASA-TLX

scores indicating that the changes made to the perturbation ratio successfully balanced subjective workload across the speed only conditions.

The behavioral and survey data indicate that our paradigm successfully achieved our goals. More importantly, in achieving the overt objectives of the study, this paradigm provides a valid start to future analysis beginning with the baseline understanding that the data were collected in accord with key constraints required for operational relevance. If our initial high level results indicated that, for instance, workload was not controlled adequately, any subsequently observed significant differences in psychophysiological variables could not be clearly attributed to TiA alone. However, more than understanding the changes in the psychophysiological variables associated with dynamic levels of TiA, is the inquiry into how changes in these variables might reflect the psychophysiological underpinning for the observed behavior, i.e., the interactions with automations, and the development of TiA. These interaction behaviors result from decisions made against a background of current psychological state which has been shown to significantly affect decision making.

Operational neuroscience studies in the context of driving automation might be aimed at understanding the psychophysiological events that support these interaction decisions, such as specific EEG and GSR features. Cognitive neuroscience research into decision making has discovered some of the neural dynamics involved in decision making. For example, fMRI studies have shown that the amygdala and the ventral striatum act to assess the valence of stimuli and that these signals are compared in the intraparietal region [17]. While operational studies necessarily use EEG rather than fMRI, these findings provide a basis for hypotheses about the cortical sources, which can be identified through localization algorithms. EEG studies aimed at understanding the cortical dynamics of complex real world decisions have identified specific frequency changes over the medial frontal regions [18]. Accompanying these neural correlates of decision making are changes in peripheral physiological and eye movement behavior. In particular, during difficult decisions, average tonic GSR magnitude increases more than if the decision was easy [19]. Eye movement, specifically gaze fixation behavior has also been associated with the cognitive processing of stimuli prior to a decision [20]. Clearly, results from cognitive neuroscience studies of decision making are fertile ground from which to generate hypotheses for further analyses in well-conducted operational experiments. We believe that our paradigm provides a research environment capable of addressing such questions.

7 Conclusion

Poor human automation integration due to mis-calibrated levels of TiA motivated an attempt to create an experimental paradigm suited to measure TiA in an operational context so that it is applicable to the real world. Because trust is a psychological state, we considered that the application of cognitive neuroscience methods rooted in psychophysiology, would be an appropriate approach to developing indicators of TiA. Typically, these methods are not used for operational neuroscience and therefore a new experimental paradigm was required. We determined through literature review what

constraints were needed for successful paradigm development. We found that if indicators of TiA were to be developed that (1) there needed to be a sense of risk or consequence, (2) that there needed to be different reliabilities of the presented automations in order to manipulate TiA levels, and (3) that workload needed to be balanced across conditions. Driving was considered to provide an optimal operational environment for our research because most adults experience driving regularly and therefore would require little training. In particular, as driving automations are becoming more common driver TiA is critical, and in the driving environment, subjects would naturally interact with automations they are familiar with. Our initial results suggest that we met these goals and that our experimental paradigm provides a valid method of studying TiA and human automation interaction in an operational setting.

Acknowledgement. This research was supported by the Office of the Secretary of Defense Autonomy Research Pilot Initiative program MIPR DWAM31168, and in part by an appointment to the U.S. Army Research Postdoctoral Fellowship Program administered by the Oak Ridge Associated Universities through a cooperative agreement with the U.S. Army Research Laboratory. Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911-NF-12-2-0019. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. We would also like to thank Dr. Justin Brooks and Dr. Javier Garcia for their advice.

References

1. Muir, B.M.: Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. Ergonomics **37**, 1905–1922 (1994)
2. Muir, B.M.: Operators' trust in and percentage of time spent using the automatic controllers in a supervisory process control task. Doctoral, University of Tornonto (1989)
3. Lee, J., Moray, N.: Trust, control strategies and allocation of function in human-machine systems. Ergonomics **35**, 1243–1270 (1992)
4. Prinzel, L.J., Freeman, F.G., Scerbo, M.W., Mikulka, P.J., Pope, A.T.: A closed-loop system for examining psychophysiological measures for adaptive task allocation. Int. J. Aviat. Psychol. **10**, 393–410 (2000)
5. Scerbo, M.: Adaptive automation. In: Neuroergonomics: The Brain at Work, pp. 239–252 (2006)
6. Borum, R.: The science of interpersonal trust (2010). Corritore, L., Kracher, B., Wiedenbeck, S.: On-line trust: concepts, evolving themes, a model. Int. J. Hum.-Comput. Stud. **58**, 737–758 (2003)
7. Lee, J.D., Moray, N.: Trust, self-confidence, and operators' adaptation automation. Int. J. Hum.-Comput. Stud. **40**, 153–184 (1994)
8. Lee, J.D., See, K.A.: Trust in automation: designing for appropriate reliance. Hum. Factors: J. Hum. Factors Ergon. Soc. **46**, 50–80 (2004)
9. Muir, B.M., Moray, N.: Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. Ergonomics **39**, 31 (1996)

10. Merritt, S.M., Ilgen, D.R.: Not all trust is created equal: dispositional and history-based trust in human-automation interactions. Hum. Factors: J. Hum. Factors Ergon. Soc. **50**, 194–210 (2008)
11. Wickens, C.D., Dixon, S.R.: The benefits of imperfect diagnostic automation: a synthesis of the literature. Theor. Issues Ergon. Sci. **8**, 201–212 (2007)
12. Parasuraman, R., Sheridan, T.B., Wickens, C.D.: A model for types and levels of human interaction with automation. IEEE Trans. Syst. Man Cybern. **30**, 10 (2000)
13. Dzindolet, M.T., Pierce, L.G., Beck, H.P., Dawe, L.A.: Misuse and disuse of automated aids. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, p. 339 (1999)
14. Wickens, C.D.: Imperfect and unreliable automation and its implications for attention allocation, information access and situation awareness (2000)
15. Cummings, M.L., Mastracchio, C., Thornburg, K.M., Mkrtchyan, A.: Boredom and distraction in multiple unmanned vehicle supervisory control. Interact. Comput. **25**, 34–47 (2013)
16. Basten, U., Biele, G., Heekeren, H.R., Fiebach, C.J.: How the brain integrates costs and benefits during decision making. Proc. Natl. Acad. Sci. **107**, 21767–21772 (2010)
17. Davis, C.E., Hauf, J.D., Wu, D.Q., Everhart, D.E.: Brain function with complex decision making using electroencephalography. Int. J. Psychophysiol. **79**, 175–183 (2011)
18. Zhou, J., Sun, J., Chen, F., Wang, Y., Taib, R., Khawaji, A., et al.: Measurable decision making with GSR and pupillary analysis for intelligent user interface. ACM Trans. Comput. Hum. Interact. (ToCHI) **21**, 33 (2015)
19. Glaholt, M.G., Reingold, E.M.: Eye movement monitoring as a process tracing methodology in decision making research. J. Neurosci. Psychol. Econ. **4**, 125 (2011)
20. Gidlöf, K. et al.: Using eye tracking to trace a cognitive process: Gaze behaviour during decision making in a natural environment. J. Eye Mov. Res. **6**(1), 1–14 (2013)